

Ultra High Density Flash Memory

Cross Reference To Related Applications

This application is related to co-pending, commonly assigned patent application serial number _____ (Attorney Docket No. 00303.342US1), filed 10 on even date herewith, which disclosure is herein incorporated by reference.

Technical Field of the Invention

This invention relates generally to integrated circuits, and particularly to floating gate transistor structures for use in nonvolatile semiconductor memories 15 such as in flash EEPROM memory cells.

Background of the Invention

Electrically erasable and programmable read only memories (EEPROMs) are reprogrammable nonvolatile memories that are widely used in computer systems for 20 storing data both when power is supplied or removed. The typical data storage element of an EEPROM is a floating gate transistor, which is a field-effect transistor (FET) having an electrically isolated (floating) gate that controls electrical conduction between source and drain regions. Data is represented by charge stored on the floating gate and the resulting conductivity obtained between source and 25 drain regions.

Increasing the storage capacity of EEPROM memories requires a reduction in the size of the floating gate transistors and other EEPROM components in order to increase the EEPROM's density. However, memory density is typically limited by a minimum lithographic feature size (F) that is imposed by lithographic

processes used during fabrication. For example, the present generation of high density dynamic random access memories (DRAMs), which are capable of storing 256 Megabits of data, require an area of $8F^2$ per bit of data. There is a need in the art to provide even higher density memories in order to further increase storage capacity.

5

Summary of the Invention

The present invention includes an ultra high density electrically erasable and programmable read only memory (EEPROM) providing increased nonvolatile 10 storage capacity. The memory allows simultaneous erasure of multiple data bits, and is referred to as flash EEPROM. Both bulk semiconductor and semiconductor-on-insulator (SOI) embodiments are included.

In one embodiment of the invention, a memory cell includes a pillar of semiconductor material that extends outwardly from a working surface of a 15 substrate. The pillar includes source/drain and body regions and has a number of sides. More than two floating gates are included in each memory cell. Each gate is associated with a side of the pillar. A number of control gates are also included in each memory cell. Each control gate is associated with a floating gate so as to allow selective storage and retrieval of data on the floating gates. In one embodiment, the 20 control gate is capable of storing more than two charge states on its associated floating gate.

In another embodiment of the invention, a memory cell is fabricated upon a substrate. The memory cell includes a first conductivity type semiconductor pillar. The pillar has top and side surfaces and is formed upon the substrate. A first 25 source/drain region, of a second conductivity type, is formed proximal to an interface between the pillar and the substrate. A second source/drain region, of a second conductivity type, is formed in a portion of the pillar that is distal to the substrate and separate from the first source/drain region. A gate dielectric is formed on at least a portion of the side surface of the pillar. More than two floating gates are

formed, each of which is substantially adjacent to a portion of the side surface of the pillar and separated therefrom by the gate dielectric. A plurality of control gates are formed, each of which is substantially adjacent to one of the floating gates and insulated therefrom. An intergate dielectric is formed, interposed between each of 5 the substantially adjacent floating and control gates.

In another embodiment of the invention a nonvolatile memory array is fabricated upon a substrate. The memory array includes a plurality of memory cells, each memory cell having a number of floating gate transistors. The memory array also includes a plurality of first gate lines that are substantially parallel in a first 10 direction. Each first gate line interconnects ones of the control gates in ones of the memory cells. A plurality of second gate lines are also formed. The plurality of second gate lines are substantially parallel in a second direction that is substantially orthogonal to the first direction. Each second gate line interconnects ones of the control gates in ones of the memory cells. At least one first source/drain 15 interconnection line is formed at least partially within the substrate. The first source/drain interconnection line interconnects ones of the first source/drain regions of ones of the memory cells. A plurality of data lines is formed. Each data line interconnects ones of the second source/drain regions of ones of the memory cells.

In another embodiment, the present invention provides a method of forming 20 a memory array. The method includes several steps, as described below. A plurality of first conductivity type semiconductor pillars are formed upon a substrate. Each pillar has top and side surfaces. A plurality of first source/drain regions, of a second conductivity type, are formed. Each of the first source/drain regions is formed proximally to an interface between the pillar and the substrate. A 25 plurality of second source/drain regions, of a second conductivity type, is formed. Each of the second source/drain regions is formed within one of the pillars and distal to the substrate and separate from the first/source drain region. A gate dielectric is formed on at least a portion of the side surfaces of the pillars. A plurality of floating gates is formed. Each of the floating gates is formed substantially adjacent to a

portion of the side surface of one of the pillars and separated therefrom by the gate dielectric. A plurality of control gates is formed. Each of the control gates is formed substantially adjacent to one of the floating gates and insulated therefrom. An intergate dielectric is formed, interposed between ones of the floating gates and ones of the control gates. A plurality of first gate lines is formed. The first gate lines are substantially parallel in a first direction. Each first gate line interconnects ones of the control gates. A plurality of second gate lines are formed, which are substantially parallel in a second direction that is substantially orthogonal to the first direction. Each second gate line interconnects ones of the control gates. At least one first source/drain interconnection line is formed at least partially within the substrate. The first source/drain interconnection line interconnects ones of the first source/drain regions. A plurality of data lines is formed. Each data line interconnects ones of the second/source drain regions.

In one embodiment, the method of forming a memory array on a substrate includes forming a first source/drain layer at a surface of the substrate. A semiconductor epitaxial layer is formed on the first source/drain layer. A second source/drain layer is formed at a surface of the epitaxial layer. A plurality of substantially parallel first troughs are etched, in a first direction, in the epitaxial layer. A first gate dielectric layer is formed substantially adjacent to sidewall regions of the first troughs. A first conductive layer is formed in the first troughs. A portion of the first conductive layer in the first troughs is removed, such that floating gate regions are formed along the sidewall regions therein and separated from the sidewall regions by the first gate dielectric layer. A portion of the substrate, underlying a portion of the first troughs and between the floating gate regions, is etched. A first intergate dielectric layer is formed on exposed portions of the floating gate regions in the first troughs. First gate lines are formed in the underlying etched portion of the substrate between opposing floating gate regions in the first troughs. Control gate regions are formed in the first troughs between

opposing floating gate regions and separated therefrom by the first intergate dielectric layer.

A plurality of substantially parallel second troughs are etched, in a second direction that is substantially orthogonal to the first direction, in the epitaxial layer.

5 A second gate dielectric layer is formed substantially adjacent to sidewall regions of the second troughs. A second conductive layer is formed in the second troughs. A portion of the second conductive layer in the second troughs is removed, such that floating gate regions are formed along the sidewall regions therein and separated from the sidewall regions by the second gate dielectric layer. A second intergate dielectric layer is formed on exposed portions of the floating gate regions in the second troughs. Control gate regions and second gate lines are formed between opposing floating gate regions in the second troughs and separated from the floating gate regions in the second troughs by the second intergate dielectric layer.

10

Thus, the present invention includes bulk semiconductor and semiconductor-on-insulator embodiments of an ultra high density flash EEPROM having increased nonvolatile storage capacity. If a floating gate transistor is used to store a single bit of data, an area of only F^2 is needed per bit of data, where F is the minimum lithographic feature size. If multiple charge states (more than two) are used, an area of less than F^2 is needed per bit of data. The increased storage capacity of the flash 15 EEPROM is particularly advantageous in replacing hard disk drive data storage in computer systems. In such an application, the delicate mechanical components included in the hard disk drive are replaced by rugged, small, and durable solid-state 20 flash EEPROM packages. The flash EEPROMs provide improved performance, extended rewrite cycles, increased reliability, lower power consumption, and 25 improved portability.

Brief Description of the Drawings

In the drawings, like numerals describe substantially similar components throughout the several views.

Figure 1 is a schematic/block diagram illustrating generally an architecture of one embodiment of a nonvolatile memory, according to the teachings of the invention, including an array having a plurality of memory cells.

Figure 2 is a schematic diagram illustrating generally one embodiment of an array of memory cells according to the teachings of the invention.

Figure 3 is a perspective view illustrating generally one embodiment of a portion of an array of memory cells according to the teachings of the invention.

Figure 4 is a plan view from above of a working surface of a substrate, which illustrates one embodiment of one of a memory cell according to the teachings of the invention.

Figures 5 - 20 illustrate generally various stages of one embodiment of a method of forming an array of memory cells according to the teachings of the invention.

Figure 21 is a perspective view of a structure resulting from another embodiment of a method of forming the array of memory cells according to the invention, using semiconductor-on-insulator (SOI) techniques.

Detailed Description of the Invention

In the following detailed description, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that the embodiments may be combined, or that other embodiments may be utilized and that structural, logical and electrical changes may be made without departing from the scope of the present invention. In the following description, the terms wafer and substrate are interchangeably used to refer generally to any structure on which integrated circuits are formed, and also to such structures during various stages of integrated circuit fabrication. Both terms include doped and undoped semiconductors, epitaxial layers of a semiconductor on

DRAFTED BY DRAFTED BY

a supporting semiconductor or insulating material, combinations of such layers, as well as other such structures that are known in the art, including bulk semiconductor and semiconductor-on-insulator (SOI) substrates. In the drawings, like numerals describe substantially similar components throughout the several views. The 5 following detailed description is not to be taken in a limiting sense.

Figure 1 is a schematic/block diagram illustrating generally an architecture of one embodiment of a memory 100 according to the present invention. In the embodiment of Figure 1, memory 100 is a nonvolatile ultra high density electrically erasable and programmable read only memory (EEPROM) allowing simultaneous 10 erasure of multiple data bits, referred to as flash EEPROM. However, the invention can be applied to other semiconductor memory devices, such as static or dynamic random access memories (SRAMs and DRAMs, respectively), synchronous random access memories or other types of memories that include a matrix of selectively addressable memory cells.

15 Memory 100 includes a memory cell array 105, having memory cells therein that include floating gate transistors, as described below. Y gate decoder 110 provides a plurality of first gate lines, YG1, YG2, ..., YGN for addressing floating gate transistors in array 105, as described below. X gate decoder 115 provides a plurality of second gate lines, XG1, XG2, ..., XGN for addressing floating gate 20 transistors in array 105, as described below. Y source/drain decoder 120 provides a plurality of first source/drain interconnection lines YS1, YS2, ..., YSN, for accessing first source/drain regions of the floating gate transistors in array 105, as described below. In an embodiment in which commonly connected first source/drain interconnection lines YS1, YS2, ..., YSN are used, Y source/drain decoder 120 may 25 be omitted. X source/drain decoder 125 provides a plurality of data lines, XD1, XD2, ..., XDN for accessing second source/drain regions of the floating gate transistors in array 105, as described below. X source/drain decoder 125 also typically includes sense amplifiers and input/output (I/O) circuitry for reading, writing, and erasing data to and from array 105. In response to address signals A0-

AN that are provided on address lines 130 during read, write, and erase operations, address buffers 135 control the operation of Y gate decoder 110, X gate decoder 115, Y source/drain decoder 120, and X source/drain decoder 125. The address signals A0-AN are provided by a controller such as a microprocessor that is

5 fabricated separately or together with memory 100, or otherwise provided by any other suitable circuits. As described in detail below, the address signals A0-AN are decoded by Y gate decoder 110, X gate decoder 115, Y source/drain decoder 120, and X source/drain decoder 125 to perform reading, writing, and erasing operations on memory cells that include a number of vertical floating gate field-effect

10 transistors (FETs) formed on the sides of a semiconductor pillar on a substrate.

10 transistors (FETs) formed on the sides of a semiconductor pillar on a substrate.

Figure 2 is a schematic diagram illustrating generally one embodiment of array 105 in more detail. In Figure 2, each memory cell 205 comprises four floating gate transistors 200, e.g. four field-effect transistors (FETs), each having an electrically isolated (floating) gate that controls electrical conduction between

15 source and drain regions. The floating gate transistors 200 are arranged in cells 205, such as cells 205AA, 205BA, ..., 205NA, in a first direction, e.g. in the Y-direction of the first source/drain interconnection lines YS1, YS2, ..., YSN, and in cells such as 205AA, 205AB, ..., 205AN in a second direction, e.g. in the X-direction of the data lines, XD1, XD2, ..., XDN. In the embodiment of Figure 2, each cell 205

20 includes four floating gate transistors 200 that share a common first source/drain region, such as a source region coupled to one of the first source/drain interconnection lines YS1, YS2, ..., YSN. The floating gate transistors 200 of each cell 205 also share a common second source/drain region, such as a drain region coupled to one of the data lines, XD1, XD2, ..., XDN. Each cell 205 has first and

25 second source/drain regions that are fabricated using a common semiconductor pillar on a substrate, as explained below.

Figure 3 is a perspective view illustrating generally one embodiment of a portion of array 105, including portions of two cells 205 of floating gate transistors 200, such as illustrated in Figure 2. In Figure 3, the substantially identical cells 205

are illustrated by way of example through cells 205AA and 205BA. Cells 205AA and 205BA each include a semiconductor pillar 300, initially of a first conductivity type such as P- silicon, fabricated upon a monolithic substrate 305. In one embodiment, substrate 305 is a bulk semiconductor, such as P- silicon. In another embodiment, a semiconductor-on-insulator (SOI) substrate 305 includes an insulating layer, such as silicon dioxide (SiO_2), as described below.

5 Each pillar 300 includes a first source/drain region of a second conductivity type, such as N⁺ silicon source region 310, formed proximally to a sub-micron dimensioned interface between pillar 300 and substrate 305. Each pillar 300 also includes a second source/drain region of the second conductivity type, such as N⁺ silicon drain region 315, that is distal to substrate 305, and separated from source region 310 by a first conductivity type region, such as P- body region 320.

10 Each pillar 300 provides a source region 310, a drain region 315, and a body region 320 for the four floating gate transistors 200 of a particular memory cell 205. In one embodiment, the physical dimensions of each pillar 300 and the doping of P- body region 320 are both sufficiently small to allow operation of the floating gate transistors 200 that is characteristic of fully depleted body transistors. First source/drain region interconnection line YS1 electrically interconnects the source region 310 of each pillar 300 of cells 205AA, 205BA, ..., 205BN. In one embodiment, the first source/drain interconnection lines YS1, YS2, .., YSN, comprise a conductively doped semiconductor of the second conductivity type, such as N⁺ silicon, disposed at least partially within substrate 305. For example, dopants can be ion-implanted or diffused into substrate 305 to form the first source/drain interconnection lines YS1, YS2, ..., YSN. In another embodiment, the first source/drain interconnection lines YS1, YS2, ..., YSN are formed above substrate 305. For example, a doped epitaxial semiconductor layer can be grown on substrate 305, from which first source/drain interconnection lines YS1, YS2, .., YSN are formed. Alternatively, an undoped epitaxial semiconductor layer can be grown on substrate 305, and dopants then introduced by ion-implantation or diffusion to

obtain the first source/drain interconnection lines **YS1**, **YS2**, ..., **YSN** of the desired conductivity.

Each pillar **300** is outwardly formed from substrate **305**, and is illustrated in Figure 3 as extending vertically upward from substrate **305**. Each pillar **300** has a

5 top region that is separated from substrate **305** by four surrounding side regions. A floating gate **325** is formed substantially adjacent to each side surface of pillar **300**, and separated therefrom by a gate dielectric **330**, such that there are four floating gates **325** per pillar **300**, though Figure 3 omits some of the floating gates **325** for clarity of illustration. Each floating gate **325** has a corresponding substantially

10 adjacent control gate **335**, from which it is separated by an intergate dielectric **340**. Except at the periphery of array **105**, each control gate **335** is interposed between two approximately adjacent pillars **300** and shared by two floating gate transistors **200**, each of these floating gate transistors **200** having portions in one of the two approximately adjacent pillars **300**.

15 Also interposed between approximately adjacent pillars **300**, except at the periphery of array **105**, are first gate lines **YG1**, **YG2**, ..., **YGN** that are substantially parallel to each other in the first direction, e.g. the Y-direction. Each of the first gate lines **YG1**, **YG2**, ..., **YGN** interconnects ones of the control gates **335**. For example, first gate line **YG1** electrically interconnects control gates **335** of floating

20 gate transistors **200** in cells **205AA**, **205BA**, ..., **205BN**. In the embodiment of Figure 3, the first gate lines **YG1**, **YG2**, ..., **YGN** are disposed at least partially within substrate **305**, as described below.

Also interposed between approximately adjacent pillars **300**, except at the periphery of array **105**, are second gate lines **XG1**, **XG2**, ..., **XGN** that are

25 substantially parallel to each other in the second direction, e.g. the X-direction. Each of the second gate lines **XG1**, **XG2**, ..., **XGN** interconnects ones of the control gates **335**. For example, second gate line **XG2** electrically interconnects control gates **335** of floating gate transistors **200**, in which the control gates are shared between pairs of cells **205**, e.g. **205AA** and **205BA**, **205AB** and **205BB**, ..., **205AN**

and 205BN. In the embodiment of Figure 3, the second gate lines XG1, XG2, ..., XGN are disposed above substrate 305, as described below.

5 Drain regions 315 of the pillars 300 are interconnected by data lines XD1, XD2, ..., XDN that are substantially parallel to each other in the second direction, e.g. the X-direction. Figure 3 illustrates, by way of example, data lines XD1 and XD2, which are shown schematically for clarity. However, it is understood that data lines XD1, XD2, ..., XDN comprise metal or other interconnection lines that are isolated from the underlying topology, e.g. pillars 300, floating gates 325, control gates 335, first gate lines YG1, YG2, ..., YGN, and second gate lines XG1, XG2, ..., XGN, by an insulating layer through which contact holes are etched to access the drain regions 315 of the pillars 300.

10 Figure 4 is a plan view, looking toward the working surface of substrate 305, illustrating generally by way of example one embodiment of one of cells 205 of four floating gate transistors 200, such as cell 205BB. In Figure 4, each of the four floating gates 325 is adjacent to one side of pillar 300, and separated therefrom by gate dielectric 330. Each control gate 335 is separated from a corresponding floating gate 325 by an intergate dielectric 340, and is integrally formed together with one of the first gate lines YG1, YG2, ..., YGN or second gate lines XG1, XG2, ..., XGN. The control gates 335 that are integrally formed together with ones of the first gate lines YG1, YG2, ..., YGN protrude upwardly therefrom such that an overlap capacitance is created with floating gates 325 that are disposed on either side thereof.

15 The center-to-center spacing (“pitch”) between adjacent first gate lines YG1, YG2, ..., YGN, such as between YG2 and YG3, or between adjacent second gate lines XG1, XG2, ..., XGN, such as between XG2 and XG3, is twice the minimum lithographic feature size F. Since four floating gate transistors 200 are contained within a cell 205 having an area of $4F^2$, an area of only F^2 is needed per bit of data. In another embodiment, multiple charge states (more than two) are used to obtain correspondingly higher data storage densities, such that an area of less than F^2 is

needed per bit of data, since more than one bit of data can be stored on a single floating gate transistor 200. In one embodiment, four charge states are used to store two bits of data per floating gate transistor 200, corresponding to eight bits of data per memory cell 205. One example of using more than two charge states to store 5 more than one bit of data per transistor is set forth in an article by T.-S. Jung et al., entitled "A 117-mm² 3.3-V Only 128-Mb Multilevel NAND Flash Memory For Mass Storage Applications," IEEE J. Solid-State Circuits, Vol. 31, No. 11, November 1996. In a further embodiment, a continuum of charge states is used to store analog data in array 105.

10 In one embodiment, programming of one of the floating gate transistors 200 is by hot electron injection. For example, a voltage of approximately 10 volts is provided, such as by one of Y gate decoder 110 or X gate decoder 115, through a particular one of the first gate lines YG1, YG2, ..., YGN or second gate lines XG1, XG2, ..., XGN to a particular control gate 335. A resulting inversion region 15 (channel) is formed in the body region 320 at the surface that is approximately adjacent to the particular one of the first gate lines YG1, YG2, ..., YGN or second gate lines XG1, XG2, ..., XGN. A voltage of approximately 5 Volts is provided, such as by X source/drain decoder 125, through a particular one of data lines XD1, XD2, ..., XDN to a particular drain region 315. A voltage of approximately 0 Volts 20 is provided, such as by Y source/drain decoder 120, through a particular one of first source/drain interconnection lines YS1, YS2, ..., YSN, to the particular source region 310 of the floating gate transistor 200. Electrons are injected onto the floating gate 325 interposed between the control gate 335 and the pillar 300 in which the particular drain region 315 is disposed. The exact value of the voltages 25 provided to the particular control gate 335 and drain region 315 will depend on the physical dimension of the floating gate transistor 200, including the thickness of the gate dielectric 330, the thickness of the intergate dielectric 340, and the separation between source region 310 and drain region 315. Alternatively, if higher voltages are provided to control gate 335, and the gate dielectric 330 and intergate dielectric

340 are made thinner, the floating gate transistor 200 may be programmed instead by Fowler-Nordheim tunneling of electrons from the body region 320, source region 310, or drain region 315.

Addressing a particular memory cell 205 for reading data includes selecting 5 a particular one of data lines XD1, XD2, ..., XDN and also selecting a particular one of first source/drain interconnection lines YS1, YS2, ..., YSN. Addressing a particular floating gate transistor 200 within the particular memory cell 205 for reading data further includes selecting a particular one of first gate lines YG1, YG2, ..., YGN or second gate lines XG1, XG2, ..., XGN.

10 In one embodiment, reading data stored on a particular floating gate transistor 200 includes providing a voltage of approximately 5. volts, such as by one of Y gate decoder 110 or X gate decoder 115, through a particular one of the first gate lines YG1, YG2, ..., YGN or second gate lines XG1, XG2, ..., XGN to the particular control gate 335 of the floating gate transistor 200. A voltage of 15 approximately 0 Volts is provided, such as by Y source/drain decoder 120, through a particular one of first source/drain interconnection lines YS1, YS2, ..., YSN, to the particular source region 310 of the particular floating gate transistor 200. A particular one of data lines XD1, XD2, ..., XDN that is switchably coupled to the drain region 315 of the floating gate transistor 200 is precharged to a positive 20 voltage by a sense amplifier in X source/drain decoder 125, then coupled to the drain region 315 to determine the conductivity state of the floating gate transistor 200 between its source region 310 and drain region 315.

If there are no electrons stored on the floating gate 325, the floating gate transistor 200 will conduct between its source region 310 and drain region 315, 25 decreasing the voltage of the particular one of data lines XD1, XD2, ..., XDN toward that voltage of its source region 310, e.g. toward a "low" binary logic level of approximately 0 Volts. If there are electrons stored on the floating gate 325, the floating gate transistor 200 will not conduct between its source region 310 and drain region 315. As a result, the sense amplifier will tend to increase the voltage of the

particular one of data lines **XD1**, **XD2**, ..., **XDN** toward a positive voltage, e.g. toward a "high" binary logic voltage level.

In one embodiment, erasure of floating gate transistors **200** includes providing an erasure voltage difference of approximately between -10 and -12 Volts

5 from a source region **310** to a corresponding control gate **335**. For example, a voltage of approximately 0 Volts is provided, such as by Y source/drain decoder **120**, to source regions **310** of floating gate transistors **200** that are interconnected by one or several first source/drain interconnection lines **YS1**, **YS2**, ..., **YSN**. A voltage of approximately between -10 and -12 Volts is provided, such as by one of

10 Y gate decoder **110** or X gate decoder **115**, through a corresponding one or several of the first gate lines **YG1**, **YG2**, ..., **YGN** or second gate lines **XG1**, **XG2**, ..., **XGN** to the control gates **335** of the floating gate transistors **200** to be erased. As a result of the negative voltage applied to the control gates **335**, electrons are removed from the corresponding floating gates **325** by Fowler-Nordheim tunneling, thereby

15 erasing the data from ones of the floating gate transistors **200**. In another example, a voltage of approximately between -5 and -6 Volts is applied to the control gates **335** and a voltage of approximately between +5 and +6 Volts is applied to the source regions **310** in order to obtain the erasure voltage difference of approximately between -10 and -12 Volts from a source region **310** to a corresponding control gate

20 **335**. The exact value of the erasure voltage difference will vary depending upon the physical dimensions of the floating gate transistor **200** and the thicknesses of gate dielectric **330** and intergate dielectric **340**.

In one embodiment, the entire array **105** of floating gate transistors **200** is simultaneously erased by applying approximately between -10 and -12 Volts to each

25 of first gate lines **YG1**, **YG2**, ..., **YGN** and second gate lines **XG1**, **XG2**, ..., **XGN**, and also applying 0 Volts to each of first source/drain interconnection lines **YS1**, **YS2**, ..., **YSN**. In another embodiment, one or more sectors of array **105** are simultaneously erased by selectively applying approximately between -10 and -12 Volts to one or more of first gate lines **YG1**, **YG2**, ..., **YGN** or second gate lines

XG1, XG2, ..., XGN, and also applying 0 Volts to one or more of first source/drain interconnection lines YS1, YS2, ..., YSN.

Figures 5 - 20 illustrate generally one embodiment of a method of forming memory array 105. In this embodiment, the array 105 is formed using bulk silicon processing techniques and is described, by way of example, with respect to a particular technology having a minimum feature size F, which is also sometimes referred to as a critical dimension (CD), of 0.4 microns. However, the process steps described below can be scaled accordingly for other minimum feature sizes without departing from the scope of the invention.

10 In Figure 5, a P- silicon starting material is used for substrate 305. A first source/drain layer 500, of approximate thickness between 0.2 microns and 0.5 microns, is formed at a working surface of substrate 305. In one embodiment, first source/drain layer 500 is N+ silicon formed by ion-implantation of donor dopants into substrate 305. In another embodiment, first source/drain layer 500 is N+ silicon formed by epitaxial growth of silicon upon substrate 305. On the first source/drain layer 500, a semiconductor epitaxial layer 505, such as P- silicon of 0.6 micron approximate thickness, is formed, such as by epitaxial growth. A second source/drain layer 510, such as N+ silicon of 150 nanometer approximate thickness, is formed at a surface of the epitaxial layer 505, such as by ion-implantation of 15 donor dopants into P- epitaxial layer 505 or by epitaxial growth of N+ silicon on P- epitaxial layer 505. A thin layer of silicon dioxide (SiO₂), referred to as pad oxide 515, is deposited on the second source/drain layer 510. Pad oxide 515 has a thickness of approximately 10 nanometers. A layer of silicon nitride (Si₃N₄), referred to as pad nitride 520, is deposited on the pad oxide 515. Pad nitride 520 has 20 a thickness of approximately 200 nanometers.

25 In Figure 6, photoresist masking and selective etching techniques are used to form, in the first direction (e.g., the Y direction, which is perpendicular to the plane of the drawing of Figure 6), a plurality of substantially parallel first troughs 600 that extend through the pad nitride 520, pad oxide 515, second source/drain layer 510,

the underlying portion of epitaxial layer 505, and at least partially into first source/drain layer 500. The photoresist is then removed.

In Figure 7, a thin silicon nitride oxidation barrier layer 700 is deposited by chemical vapor deposition (CVD) to protect against oxidation of sidewalls of first troughs 600. Barrier layer 700 is anisotropically etched to expose bottom portions of first troughs 600. A bottom insulation layer 705 of silicon dioxide is formed on the bottoms of first troughs 600 by thermal oxidation of the exposed bottom portions of first troughs 600.

In Figure 8, barrier layer 700 is stripped from the sidewalls of the first troughs 600, such as by a brief phosphoric acid etch, which is timed to expose the sidewalls of the first troughs 600 but which avoids significant removal of the pad nitride 520. A first gate dielectric layer 800 such as, for example, silicon dioxide of thickness approximately between 5 nanometers and 10 nanometers (sometimes referred to as “tunnel oxide”), is formed substantially adjacent to the exposed sidewalls of the first troughs 600. A first conductive layer 805, such as N+ doped polysilicon, is formed in the first troughs 600, such as by CVD, to fill the first troughs 600. The first conductive layer 805 is planarized, such as by chemical mechanical polishing (CMP) or other suitable planarization technique.

In Figure 9, the first conductive layer 805 is etched back in the first troughs 600 to approximately 100 nanometers below the silicon surface, which is defined by the interface between the second source/drain layer 510 and the pad oxide 515 layer. A first spacer layer, such as silicon nitride of an approximate thickness of 7 nanometers, is deposited by CVD and anisotropically etched by reactive ion etching (RIE) to leave nitride first spacers 900 along the sidewalls of the first troughs 600. A second spacer layer, such as silicon dioxide of an approximate thickness of 90 nanometers, is deposited by CVD and anisotropically etched by RIE to leave second spacers 905 along the sidewalls of the first troughs 600.

In Figure 10, a portion of the first conductive layer 805 in first troughs 600 between second spacers 905 is removed, such as by using spacers 905 as a mask

while etching down to bottom insulation layer 705, thereby forming from the first conductive layer 805 floating gate regions 1000 along the sidewalls of the first troughs 600. A thin oxidation barrier layer 1005, such as silicon nitride of approximate thickness of 5 nanometers, is deposited by CVD. Barrier layer 1005 is 5 removed from the bottom insulation layer 705 in first troughs 600 by anisotropic etching. The remaining portions of barrier layer 1005 protect the floating gate regions 1000 during subsequent processing described below.

In Figure 11, a portion of the bottom insulation layer 705 is removed, exposing a portion of the underlying substrate 305, by an anisotropic etch that is 10 timed to leave enough of second spacers 905 to protect floating gate regions 1000 during a subsequent etch of substrate 305. A portion of substrate 305 that underlies a portion of first troughs 600 between the floating gate regions 1000 is removed by selectively anisotropically etching the substrate 305 to a depth sufficient to carry the first gate lines YG1, YG2, ..., YGN. A first trough insulation layer 1100 is formed 15 on sidewall and bottom regions of the etched portions of substrate 305 underlying the first troughs 600. Barrier layer 1005 is removed to expose the floating gate regions 1000 in first troughs 600, such as by wet etching.

The first intergate dielectric 340, having an approximate thickness between 7 nanometers and 15 nanometers, is formed on the exposed portions of floating gate 20 regions 1000. In one embodiment, a silicon dioxide intergate dielectric 340 is formed by thermal oxidation of the floating gate regions 1000. In another embodiment, an oxynitride intergate dielectric 340 is formed on the floating gate regions 1000 by CVD.

First gate lines YG1, YG2, ..., YGN are formed in the etched portions of 25 substrate 305 underlying the first troughs 600 between opposing floating gate regions 1000 in the first troughs 600. First gate lines YG1, YG2, ..., YGN are insulated from substrate 305 by first trough insulation layer 1100. Control gates 335 are formed in the first troughs 600 between opposing floating gate regions 1000, and separated therefrom by the first intergate dielectric 340. In one embodiment, first

gate lines **YG1**, **YG2**, ..., **YGN** and control gates **335** are formed together by depositing **N+** polysilicon to fill first troughs **600**, and etching back the deposited **N+** polysilicon approximately to the top portion of the floating gate regions **1000**.

In Figure 12, a cap layer **1200** is formed, such as by CVD of silicon dioxide, 5 and then planarized, such as by CMP, such that the top surface of cap layer **1200** is substantially even with the top surface of pad nitride **520**. A masking layer **1205** is formed, such as silicon nitride deposited by CVD to an approximate thickness of 100 nanometers. Another masking layer **1210** is also formed, such as polysilicon deposited by CVD to an approximate thickness of 100 nanometers. A photoresist 10 layer **1215** is formed on masking layer **1210**.

Figure 13 is a perspective view, illustrating the selective etching, in a second direction (X-direction) that is substantially orthogonal to the first direction (Y-direction), of a plurality of substantially parallel second troughs **1300**, as described below. Forming second troughs **1300** includes selectively etching masking layer 15 **1210** and underlying masking layer **1205**, such that portions of cap layer **1200** in the second troughs **1300** are exposed. With photoresist layer **1215** still in place, a nonselective dry etch is used to simultaneously remove exposed silicon dioxide and polysilicon in intersecting portions of first troughs **600** and second troughs **1300**, including the removing of: portions of cap layer **1200**, gate dielectric **800**, floating 20 gate regions **1000**, intergate dielectric **340**, and the control gate **335** portions of first gate lines **YG1**, **YG2**, ..., **YGN**. The nonselective dry etch removal proceeds at least to the depth of the interface between floating gate regions **1000** and underlying bottom insulation layer **705**, thereby separating floating gate regions **1000** into the isolated floating gates **325**. During the nonselective dry etch, the regions between 25 first troughs **600** are protected by the pad nitride **520** and the regions between second troughs **1300** are protected by selectively patterned photoresist layer **1215**.

In the plan view of Figure 14, the photoresist layer **1215** has been removed by conventional photoresist stripping techniques, thereby exposing the underlying, selectively patterned polysilicon masking layer **1210**. An insulating layer **1400**,

such as silicon dioxide deposited by CVD, is formed everywhere on the topography of the working surface of substrate 305, thereby filling the nonselectively dry-etched intersections of the first troughs 600 and second troughs 1300. The insulating layer 1400 is then planarized, such as by CMP, and recess etched to a depth that is 5 slightly above the interface between second source-drain layer 510 and pad oxide 515, thereby leaving behind recessed portions of insulating layer 1400 in the nonselectively dry-etched intersections of the first troughs 600 and second troughs 1300, as illustrated in Figure 14.

In the plan view of Figure 15, the exposed portions of pad nitride 520 (e.g., 10 between first troughs 600 and within second troughs 1300) are removed by a selective etch of silicon nitride, thereby exposing underlying portions of pad oxide 515. The exposed portions of pad oxide 515 (e.g., between first troughs 600 and within second troughs 1300) are removed by dipping into a wet etchant, which is timed to remove the exposed portions of pad oxide 515, but to leave most of the 15 remaining portions of the thicker silicon dioxide insulating layer 1400 intact. The removing of portions of pad oxide 515 exposes the second source/drain layer 510 portion of the underlying silicon epitaxial layer 505. The exposed portions of silicon epitaxial layer 505 (e.g., between first troughs 600 and within second troughs 1300) are removed by a selective etching that is preferential to silicon over 20 silicon dioxide, thereby forming recesses 1500 in second troughs 1300 between first troughs 600. Recesses 1500, which are considered to be part of second troughs 1300, are etched through epitaxial layer 505 and at least partially into first source/drain layer 500. Etching recesses 1500 also removes the remaining portions 25 of polysilicon masking layer 1210, thereby exposing underlying silicon nitride masking layer 1205, as illustrated in Figure 15.

Figure 16 is a cross-sectional view in the direction of second troughs 1300 (e.g. such that the X-direction is orthogonal to the plane of the illustration of Figure 16), as indicated by the cut line 16 - 16 in Figure 15. In Figure 16, a thin silicon nitride oxidation barrier layer 1600 is deposited by CVD to protect against oxidation

of sidewalls of second troughs 1300. Barrier layer 1600 is anisotropically etched to expose bottom portions of second troughs 1300. A bottom insulation layer 1605 of silicon dioxide is formed on the bottoms of second troughs 1300, such as silicon dioxide of approximate thickness of 50 nanometers formed by thermal oxidation of the exposed bottom portions of second troughs 1300.

In Figure 17, barrier layer 1600 is stripped from the sidewalls of the second troughs 1300, such as by a brief phosphoric acid etch, which is timed to expose the sidewalls of the second troughs 1300 but which avoids significant removal of the silicon nitride masking layer 1205. A second gate dielectric layer 1700, such as silicon dioxide of thickness approximately between 5 nanometers and 10 nanometers (sometimes referred to as “tunnel oxide”), is formed substantially adjacent to the exposed sidewalls of the second troughs 1300. A second conductive layer 1705, such as N⁺ doped polysilicon, is formed in the second troughs 1300, such as by CVD, to fill the second troughs 1300. The second conductive layer 1705 is planarized, such as by chemical mechanical polishing (CMP) or other suitable planarization technique.

In Figure 18, the second conductive layer 1705 is etched back in the second troughs 1300 to approximately at or slightly above the level of the silicon surface, which is defined by the interface between the second source/drain layer 510 and the pad oxide 515 layer. Thus, in the second troughs 1300, the top surface of the second conductive layer 1705 is approximately even with the top surface of the recessed portions of insulating layer 1400. A spacer layer, such as silicon nitride of an approximate thickness of 100 nanometers, is deposited by CVD and anisotropically etched by reactive ion etching (RIE) to leave nitride third spacers 1800 along the sidewalls of the second troughs 1300, e.g. on the etched back portions of the second conductive layer 1705 and on the recessed portions of insulating layer 1400, and against the second gate dielectric 1700.

In the perspective view of Figure 19, third spacers 1800 are used as a mask for the anisotropic etching of the etched back portions of polysilicon second

conductive layer 1705 together with the recessed portions of silicon dioxide insulating layer 1400. By first utilizing an etchant to remove silicon dioxide, the second troughs 1300 are etched in insulating layer 1400 to a depth sufficient to carry a second gate line X1, X2, ..., XN, but not so great as to expose the first gate lines 5 Y1, Y2, ..., YN underlying the recessed portions of silicon dioxide insulating layer 1400 in second troughs 1300. Then, the anisotropic etch is continued using a selective etchant to remove polysilicon but not silicon dioxide until the bottom insulation layer 1605 is exposed, thereby forming from the second conductive layer 1705 separate floating gates 325 along the sidewalls of the second troughs 1300.

10 In the perspective view of Figure 20, a second intergate dielectric 2000 is formed in the second troughs 1300, such that the second intergate dielectric 2000 has an approximate thickness between 7 nanometers and 15 nanometers and being formed by thermal growth of silicon dioxide or deposition of oxynitride by CVD. Control gates 335 are formed between opposing floating gates 325 in the second 15 troughs 1300 and separated therefrom by the second intergate dielectric 2000. The control gates 335 in second troughs 1300 are formed together with the second gate lines X1, X2, ..., XN in second troughs 1300 by a single deposition of N⁺ doped polysilicon that fills second troughs 1300 and is planarized, such as by CMP. Phosphoric acid is used to remove the remaining silicon nitride, such as third 20 spacers 1800, masking layer 1205, and pad nitride 520, leaving the structure illustrated in Figure 20. An insulator such as silicon dioxide is then deposited, and subsequent processing follows conventional techniques for forming contact holes, terminal metal, and inter level insulator steps to complete wiring of the cells 205 and other circuits of memory 100.

25 Though Figures 5 - 20 illustrate generally one embodiment of forming the memory array 105 using bulk silicon processing techniques, in another embodiment a semiconductor-on-insulator (SOI) substrate is formed from substrate 305. In one such embodiment, a P- silicon starting material is used for substrate 305, and processing proceeds similarly to the bulk semiconductor embodiment described in

Figure 5 - 7. However, after the barrier layer 700 is formed in Figure 7, an isotropic chemical etch is used to fully undercut the semiconductor regions separating the first troughs 600, and a subsequent oxidation step is used to fill in the evacuated regions formed by the undercutting. As a result, an insulator is formed on the bottoms of 5 first troughs 600, bars of SOI are formed between first troughs 600, and the topography on the working surface of substrate 305 is separated from substrate 305 by an insulating layer 2100 illustrated in the perspective view of Figure 21.

Thus, in the above described Figures, substrate 305 is understood to include 10 bulk semiconductor as well as SOI embodiments in which the semiconductor integrated circuits formed on the surface of substrate 305 are isolated from each other and an underlying semiconductor portion of substrate 305 by an insulating layer.

One such method of forming bars of SOI is described in the Noble U.S. Patent application serial number 08/745,708 which is assigned to the assignee of the 15 present application and which is herein incorporated by reference. Another such method of forming regions of SOI is described in the Forbes U.S. Patent application serial number 08/706,230, which is assigned to the assignee of the present application and which is herein incorporated by reference.

In an SOI embodiment of the present invention, processing of first troughs 20 600 to carry the first gate lines YG1, YG2, ..., YGN varies slightly from the bulk semiconductor embodiment described with respect to Figures 10 and 11. A barrier layer 1005 need not be formed to protect the floating gate regions 1000. A portion of the substrate 305 that underlies a portion of the first troughs 600 between the floating gate regions 1000 is removed by selectively anisotropically etching the 25 silicon dioxide insulator portion of substrate 305 to a depth sufficient to carry the first gate lines YG1, YG2, ..., YGN. A portion of the resulting structure of array 105 is illustrated in the perspective view of Figure 21, which includes an insulating layer 2100 portion of substrate 305, as described above.

Thus, the present invention provides an ultra high density flash EEPROM having increased nonvolatile storage capacity. If a floating gate transistor 200 is used to store a single bit of data, an area of only F^2 is needed per bit of data. If multiple charge states (more than two) are used, an area of less than F^2 is needed per 5 bit of data. The increased storage capacity of the ultra high density flash EEPROM is particularly advantageous in replacing hard disk drive data storage in computer systems. In such an application, the delicate mechanical components included in the hard disk drive are replaced by rugged, small, and durable solid-state ultra high density flash EEPROM packages. The ultra high density flash EEPROMs provide 10 improved performance, extended rewrite cycles, increased reliability, lower power consumption, and improved portability.

It is to be understood that the above description is intended to be illustrative, and not restrictive. Many other embodiments will be apparent to those of skill in the art upon reviewing the above description. For example, though the memory cells 15 205 have been described with respect to a particular embodiment having four floating gate transistors 200 per pillar 300, a different number of floating gate transistors per pillar could also be used. It is also understood that the above structures and methods, which have been described with respect to EEPROM memory devices having floating gate transistors 200, are also applicable to dynamic 20 random access memories (DRAMs) or other integrated circuits using vertically oriented field-effect transistors (s) that do not have floating gates. Thus, the scope of the invention is not limited to the particular embodiments shown and described herein.